

Multidimensional Subset Scanning for the Public Good

Daniel B. Neill
H.J. Heinz III College
Carnegie Mellon University
E-mail: neill@cs.cmu.edu

We gratefully acknowledge funding support from the National Science Foundation, grants IIS-0916345, IIS-0911032, and IIS-0953330, and the John D. and Catherine T. MacArthur Foundation.

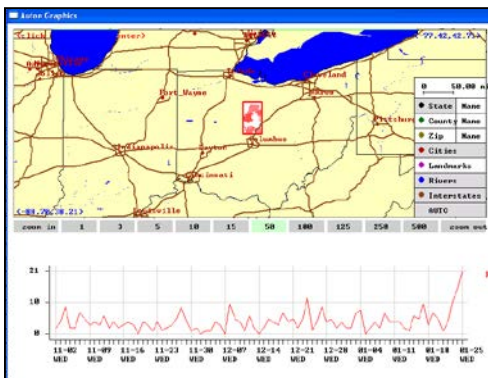
Carnegie Mellon University

EPD Lab

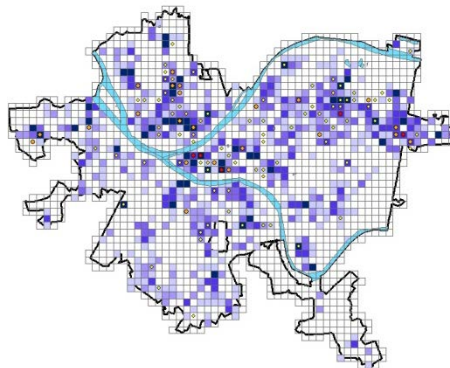
EVENT AND PATTERN DETECTION LABORATORY



Daniel B. Neill (neill@cs.cmu.edu)
Associate Professor of Information Systems, Heinz College, CMU
Director, Event and Pattern Detection Laboratory
Courtesy Associate Professor of Machine Learning and Robotics



Disease Surveillance:
Very early and accurate detection of emerging outbreaks.



Law Enforcement:
Detection, prediction, and prevention of “hot-spots” of violent crime.



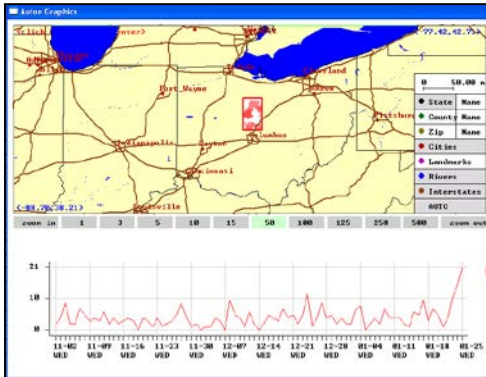
Medicine: Discovering new “best practices” of patient care, to improve outcomes and reduce costs.

My research is focused at the intersection of **machine learning** and **public policy**, with two main goals:

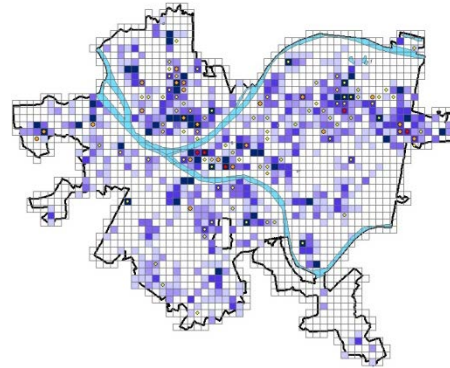
- 1) Develop new machine learning methods for better (more scalable and accurate) **detection** and **prediction** of events and other patterns in massive datasets.
- 2) Apply these methods to improve the quality of public health, safety, and security.



Daniel B. Neill (neill@cs.cmu.edu)
Associate Professor of Information Systems, Heinz College, CMU
Director, Event and Pattern Detection Laboratory
Courtesy Associate Professor of Machine Learning and Robotics



Disease Surveillance:
Very early and accurate detection of emerging outbreaks.



Law Enforcement:
Detection, prediction, and prevention of “hot-spots” of violent crime.



Medicine: Discovering new “best practices” of patient care, to improve outcomes and reduce costs.

Our disease surveillance methods have been in use for deployed systems in the U.S., Canada, India, and Sri Lanka.

Our “CrimeScan” software has been in day-to-day operational use for predictive policing by Chicago and Pittsburgh PDs. “CityScan” has been used by Chicago city leaders for prediction and prevention of rodent infestations using 311 call data.

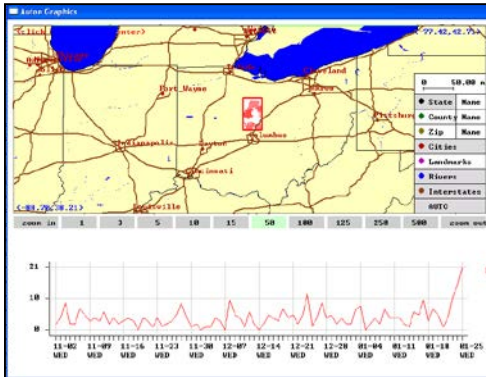


Daniel B. Neill (neill@cs.cmu.edu)

Associate Professor of Information Systems, Heinz College, CMU

Director, Event and Pattern Detection Laboratory

Courtesy Associate Professor of Machine Learning and Robotics



Disease Surveillance:

Very early and accurate detection of emerging outbreaks.

Our disease surveillance methods have been in use for deployed systems in the U.S., Canada, India, and Sri Lanka.

“CrimeScan was set up to run daily, completely autonomously. Predictions were sent to police analysts, and messages were compiled into detailed intelligence reports disseminated through the chain of command.

*Based upon deployment suggestions indicated in the CrimeScan reports, **important arrests were effected, weapons were seized, and crimes were prevented.**”*

Our “CrimeScan” software has been in day-to-day operational use for predictive policing by Chicago and Pittsburgh PDs. “CityScan” has been used by Chicago city leaders for prediction and prevention of rodent infestations using 311 call data.

Pattern detection by subset scan

One key insight that underlies much of my work is that pattern detection can be viewed as a **search** over subsets of the data.

Statistical challenges:

Which subsets to search?
Is a given subset anomalous?
Which anomalies are relevant?

Computational challenge:

How to make this search over subsets efficient for massive, complex, high-dimensional data?

New statistical methods enable more timely and more accurate detection by integrating **multiple data sources**, incorporating **spatial** and **temporal** information, and using **prior knowledge** of a domain.

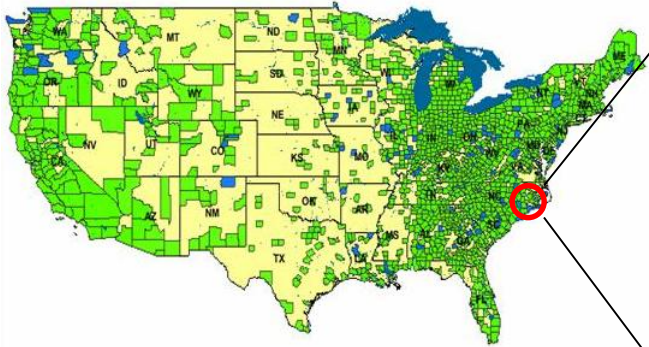
New algorithms and data structures make previously impossible detection tasks computationally feasible and fast.

New machine learning methods enable our systems to learn from user feedback, modeling and distinguishing between relevant and irrelevant types of anomaly.

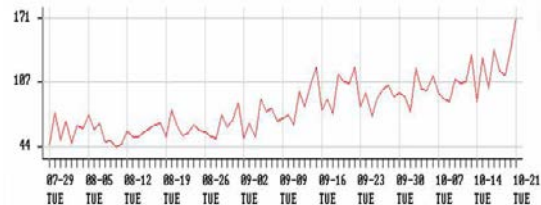
Outline of this talk

- Subset scanning for pattern detection
- Multidimensional subset scan
- Application #1: Event detection
 - outbreak detection, drug overdose surveillance
- Application #2: Discovery of heterogeneous treatment effects from observational data
 - patterns of patient care that impact outcomes
- Application #3: Auditing black-box classifiers to discover systematic biases
 - bias in criminal justice recidivism risk prediction

Multivariate event detection



Spatial time series data from spatial locations s_i (e.g. zip codes)



Time series of counts $c_{i,m}^t$ for each zip code s_i for each data stream d_m .

Outbreak detection

- d_1 = respiratory ED
- d_2 = constitutional ED
- d_3 = OTC cough/cold
- d_4 = OTC anti-fever
(etc.)

Main goals:

- Detect** any emerging events.
- Pinpoint** the affected subset of locations and time duration.
- Characterize** the event, e.g., by identifying the affected streams.

Compare hypotheses:

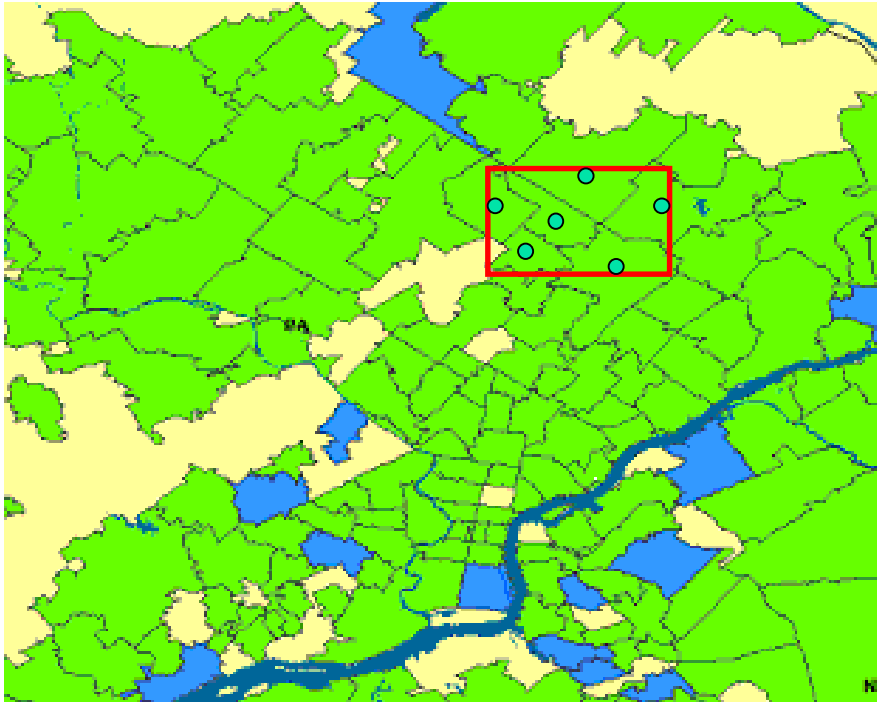
$$H_1(D, S, W)$$

- D = subset of streams
- S = subset of locations
- W = time duration

vs. H_0 : no events occurring

Expectation-based scan statistics

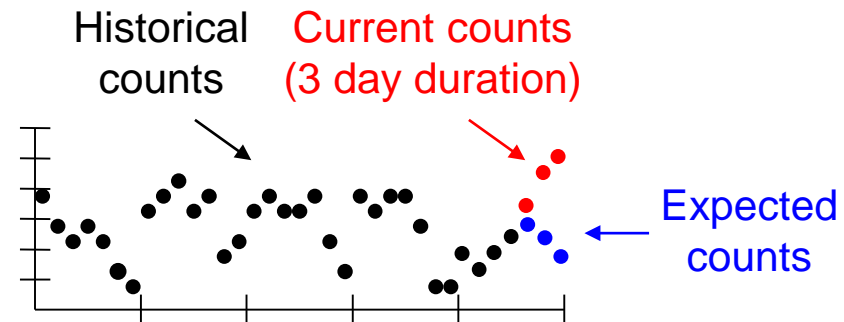
(Kulldorff, 1997; Neill and Moore, 2005)



We search for spatial regions (subsets of locations) where the recently observed counts for some subset of streams are significantly higher than expected.

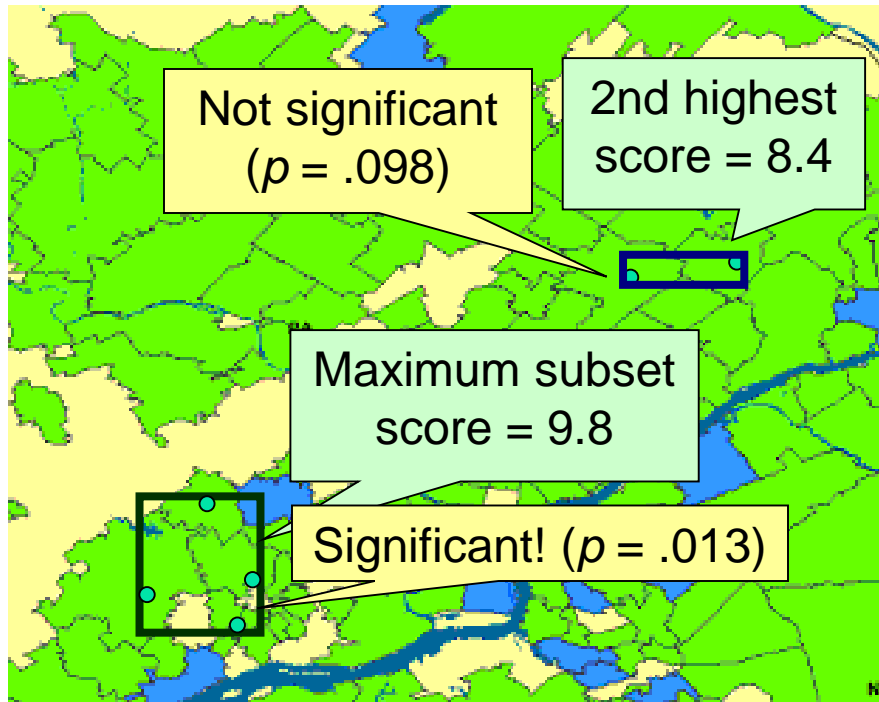
We perform **time series analysis** to compute expected counts (“baselines”) for each location and stream for each recent day.

We then compare the actual and expected counts for each subset (D, S, W) under consideration.



Expectation-based scan statistics

(Kulldorff, 1997; Neill and Moore, 2005)

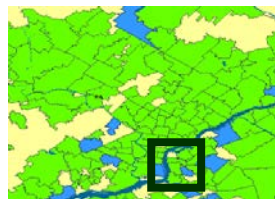


We find the subsets with highest values of a **likelihood ratio statistic**, and compute the p -value of each subset by randomization testing.

$$F(D, S, W) = \frac{\Pr(\text{Data} \mid H_1(D, S, W))}{\Pr(\text{Data} \mid H_0)}$$

To compute p -value
Compare subset score to maximum subset scores of simulated datasets under H_0 .

$$F_1^* = 2.4$$



$$F_2^* = 9.1$$



...

$$F_{999}^* = 7.0$$



Likelihood ratio statistics

For our expectation-based scan statistics, the null hypothesis H_0 assumes “business as usual”: each count $c_{i,m}^t$ is drawn from some parametric distribution with mean $b_{i,m}^t$. $H_1(S)$ assumes a multiplicative increase for the affected subset S .

Expectation-based Poisson

$$H_0: c_{i,m}^t \sim \text{Poisson}(b_{i,m}^t)$$

$$H_1(S): c_{i,m}^t \sim \text{Poisson}(qb_{i,m}^t)$$

$$\text{Let } C = \sum_S c_{i,m}^t \text{ and } B = \sum_S b_{i,m}^t.$$

$$\text{Maximum likelihood: } q = C / B.$$

$$F(S) = C \log (C/B) + B - C$$

Expectation-based Gaussian

$$H_0: c_{i,m}^t \sim \text{Gaussian}(b_{i,m}^t, \sigma_{i,m}^t)$$

$$H_1(S): c_{i,m}^t \sim \text{Gaussian}(qb_{i,m}^t, \sigma_{i,m}^t)$$

$$\text{Let } C' = \sum_S c_{i,m}^t b_{i,m}^t / (\sigma_{i,m}^t)^2 \\ \text{and } B' = \sum_S (b_{i,m}^t)^2 / (\sigma_{i,m}^t)^2.$$

$$\text{Maximum likelihood: } q = C' / B'.$$

$$F(S) = (C')^2 / 2B' + B'/2 - C'$$

Many possibilities: exponential family, nonparametric, Bayesian...

Which regions to search?

Typical approach: “spatial scan” (Kulldorff, 1997)

Each search region S is a **sub-region** of space.

- Choose some region shape (e.g. circles, rectangles) and consider all regions of that shape and varying size.
- Low power for true events that do not correspond well to the chosen set of search regions (e.g. irregular shapes).

Our approach: “subset scan” (Neill, 2012)

Each search region S is a **subset** of locations.

- Find the highest scoring subset, subject to some constraints (e.g. spatial proximity, connectivity).
- For multivariate, also optimize over subsets of streams.
- Exponentially many possible subsets, $O(2^N \times 2^M)$: computationally infeasible for naïve search.

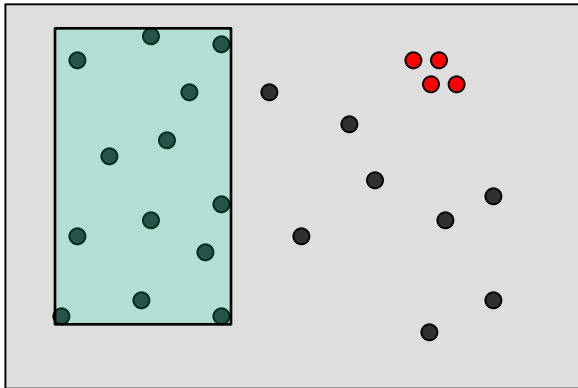
Question: Why search over subsets?

Answer: Simpler approaches can fail.

Top-down detection approaches

Are there any globally interesting patterns? If so, recursively search the most interesting sub-partition.

Two examples: bump hunting;
“cluster then detect”.

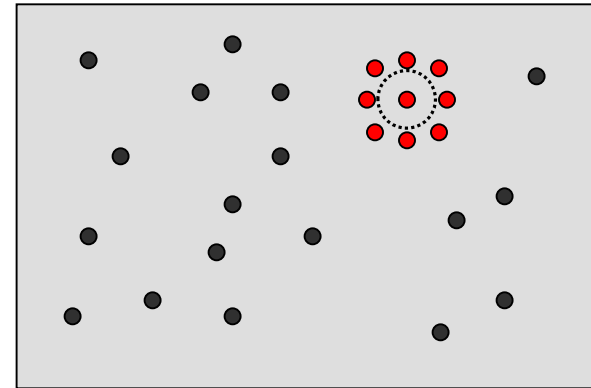


Top-down fails for **small-scale patterns** that are not evident from the global aggregates.

Bottom-up detection approaches

Find individually (or locally) anomalous data points, and optionally, aggregate into clusters.

Two examples: anomaly/outlier detection;
density-based clustering.



Bottom-up fails for **subtle patterns** that are only evident when a group of data records are considered collectively.

Question: Why search over subsets? Answer: Simpler approaches can fail.

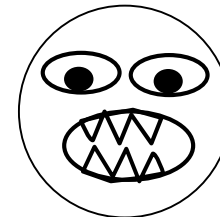
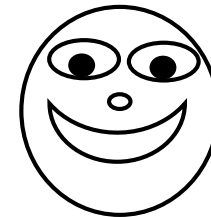
Top-down detection approaches

Are there any patterns?
the most interesting

So here's where we are so far:

Treating pattern detection as a subset scan problem is statistically desirable for maximizing detection power...

but computationally infeasible (for exhaustive search at least).



Top-down fails to find **subtle patterns** that are not evident when a group of data records are considered collectively.

Fast subset scan (Neill, 2012)

- In certain cases, we can optimize $F(S)$ over the exponentially many subsets of the data, while evaluating only $O(N)$ rather than $O(2^N)$ subsets.
- Many commonly used scan statistics have the property of linear-time subset scanning:
 - Just sort the data records (or spatial locations, etc.) from highest to lowest priority according to some function...
 - ... then search over groups consisting of the top-k highest priority records, for $k = 1..N$.

The highest scoring subset is **guaranteed** to be one of these!

Sample result: we can find the **most anomalous** subset of Allegheny County zip codes in **0.03 sec** vs. **10^{24} years**.

Linear-time subset scanning

- Example: Expectation-Based Poisson statistic
 - Sort data locations s_i by the ratio of observed to expected count, c_i / b_i .
 - Given the ordering $s_{(1)} \dots s_{(N)}$, we can **prove** that the top-scoring subset $F(S)$ consists of the locations $s_{(1)} \dots s_{(k)}$ for some k , $1 \leq k \leq N$.
 - Key step: if there exists some location $s_{\text{out}} \notin S$ with higher priority than some location $s_{\text{in}} \in S$, then we can show that $F(S) \leq \max(F(S \cup \{s_{\text{out}}\}), F(S \setminus \{s_{\text{in}}\}))$.
- Theorem: LTSS holds for expectation-based scan statistics in any exponential family. (Speakman et al., 2016)

$$F(S) = \max_{q>1} \log \frac{P(\text{Data} \mid H_1(S))}{P(\text{Data} \mid H_0)} \quad \begin{array}{l} H_0 : x_i \sim \text{Dist}(\mu_i) \\ H_1 : x_i \sim \text{Dist}(q\mu_i) \end{array}$$

Linear-time subset scanning

- Example: Expectation-Based Poisson statistic
 - Sort data locations s_i by the ratio of observed to expected count, c_i / b_i .
 - Given the ordering $s_{(1)} \dots s_{(N)}$, we can **prove** that the top-scoring subset $F(S)$ consists of the locations $s_{(1)} \dots s_{(k)}$ for some k , $1 \leq k \leq N$.
 - Key step: if there exists some location $s_{\text{out}} \notin S$ with higher priority than some location $s_{\text{in}} \in S$, then we can show that $F(S) \leq \max(F(S \cup \{s_{\text{out}}\}), F(S \setminus \{s_{\text{in}}\}))$.
- Even better theorem: We can also maximize the **penalized** scan statistic $F(S) + \sum_{s_i \in S} \Delta_i$ in $O(N \log N)$ time, evaluating only $2N$ of the 2^N subsets.

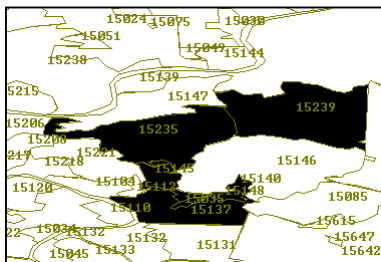
(Speakman et al., 2016)

Constrained fast subset scanning

LTSS is a new and powerful tool for **exact** combinatorial optimization (as opposed to approximate techniques such as submodular function optimization). But it only solves the “best unconstrained subset” problem, and cannot be used directly for constrained optimization.

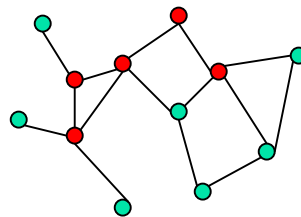
Many of our recent papers have focused on how LTSS can be extended to the many real-world problems with (hard or soft) constraints on our search.

- + Proximity constraints → Fast spatial scan (irregular regions)
- + Multiple data streams → Fast multivariate scan
- + Connectivity constraints → Fast graph scan
- + Group self-similarity → Fast generalized subset scan

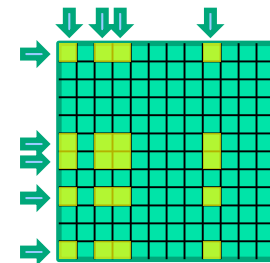


(Neill, *JRSS-B*, 2012)

(Neill et al., *Stat. Med.*, 2013)



(Speakman et al., *JCGS*, 2015)



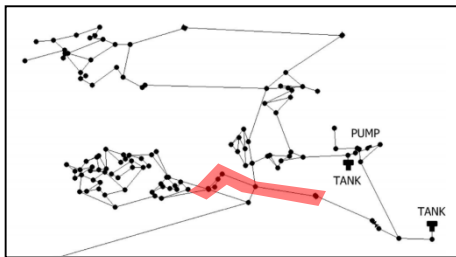
(McFowland et al., *JMLR*, 2013)

Constrained fast subset scanning

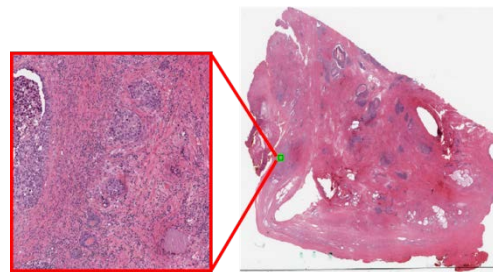
LTSS is a new and powerful tool for **exact** combinatorial optimization (as opposed to approximate techniques such as submodular function optimization). But it only solves the “best unconstrained subset” problem, and cannot be used directly for constrained optimization.

Many of our recent papers have focused on how LTSS can be extended to the many real-world problems with (hard or soft) constraints on our search.

- + Temporal dynamics → Spreading contamination in water supply
- + Hierarchical scanning → Prostate cancer in digital pathology slides
- + Scalable GP regression → Predicting and preventing rat infestations



(Speakman et al., ICDM 2013)



(Somanchi & Neill, DMHI 2013)



(Flaxman et al., 2015;
Neill et al., in preparation)

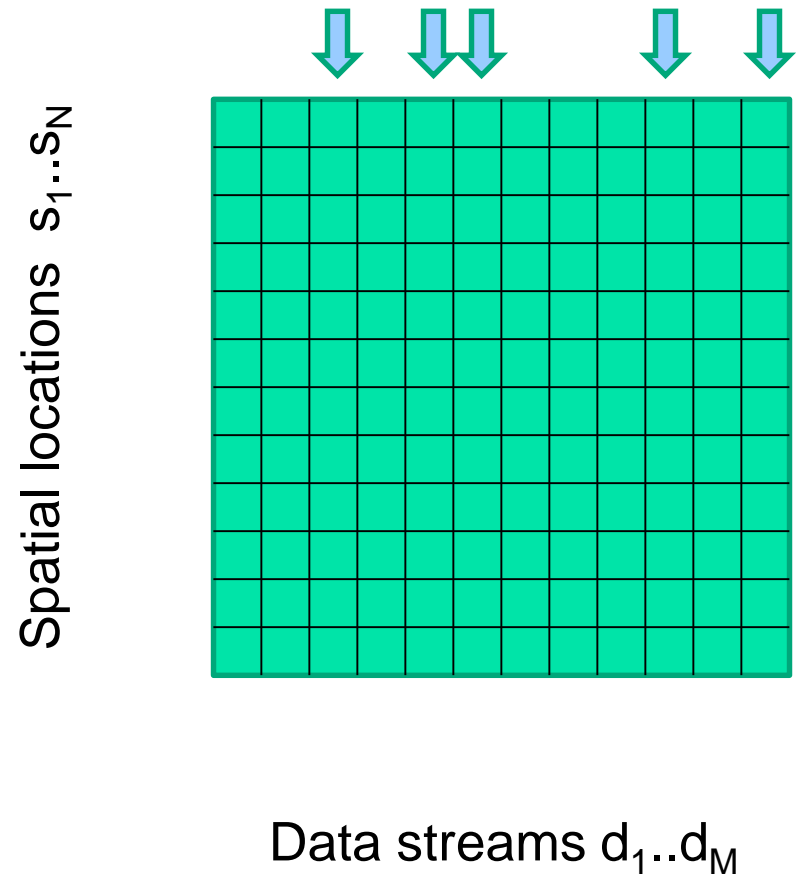
Fast subset scan with spatial proximity constraints

- Maximize a likelihood ratio statistic over all subsets of the “local neighborhoods” consisting of a center location s_i and its $k-1$ nearest neighbors, for a fixed neighborhood size k .
- Naïve search requires $O(N \cdot 2^k)$ time and is computationally infeasible for $k > 25$.
- For each center, we can search over all subsets of its local neighborhood in $O(k)$ time using LTSS, thus requiring a total time complexity of $O(Nk) + O(N \log N)$ for sorting the locations.
- In Neill (2012), we show that this approach dramatically improves the timeliness and accuracy of outbreak detection for irregularly-shaped disease clusters.

Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

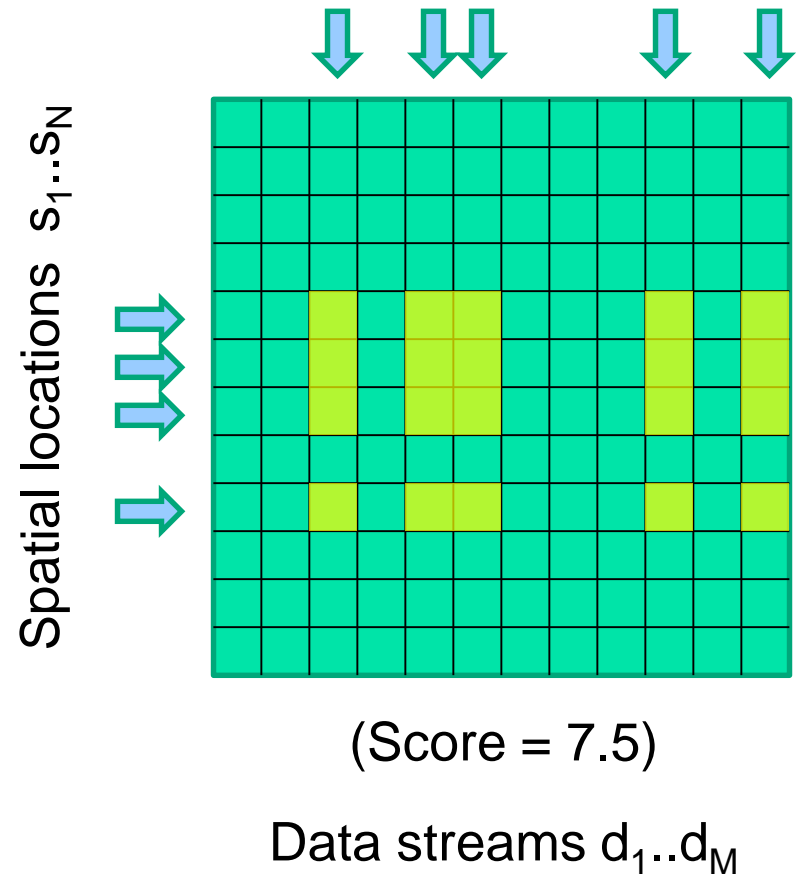
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.



Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

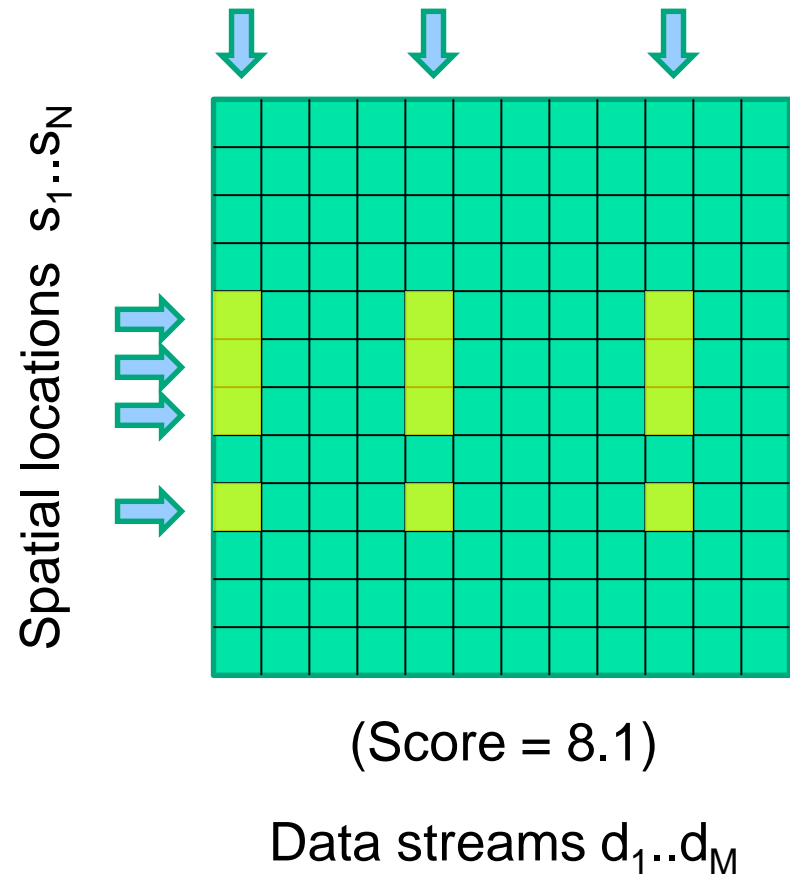
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.



Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

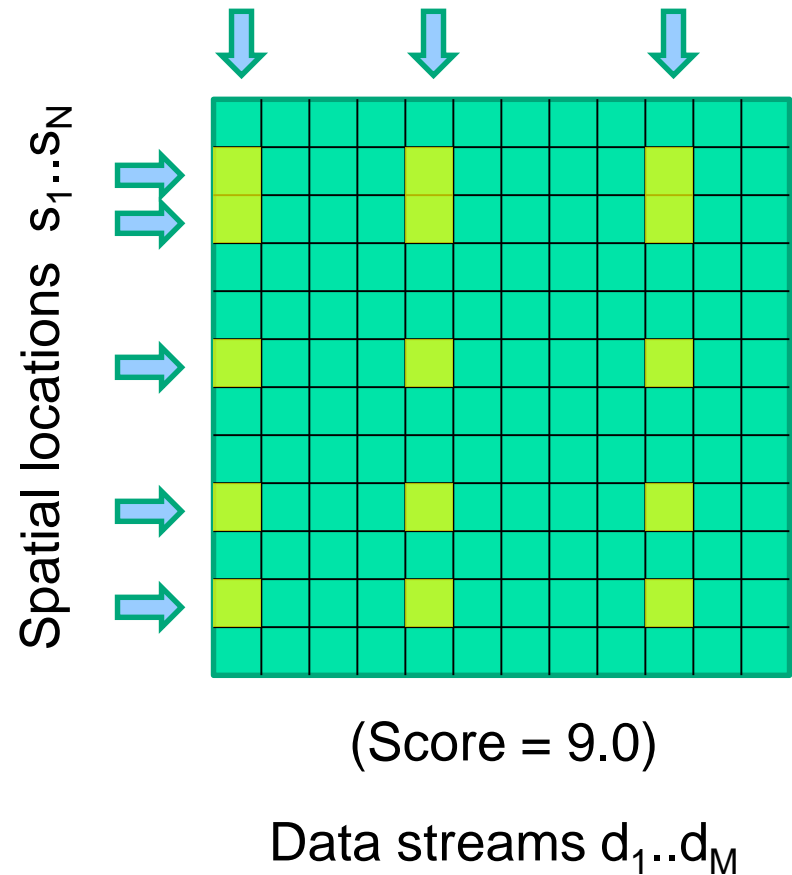
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.
- But it also allows us to efficiently optimize over subsets of **streams** for a given subset of **locations**...



Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

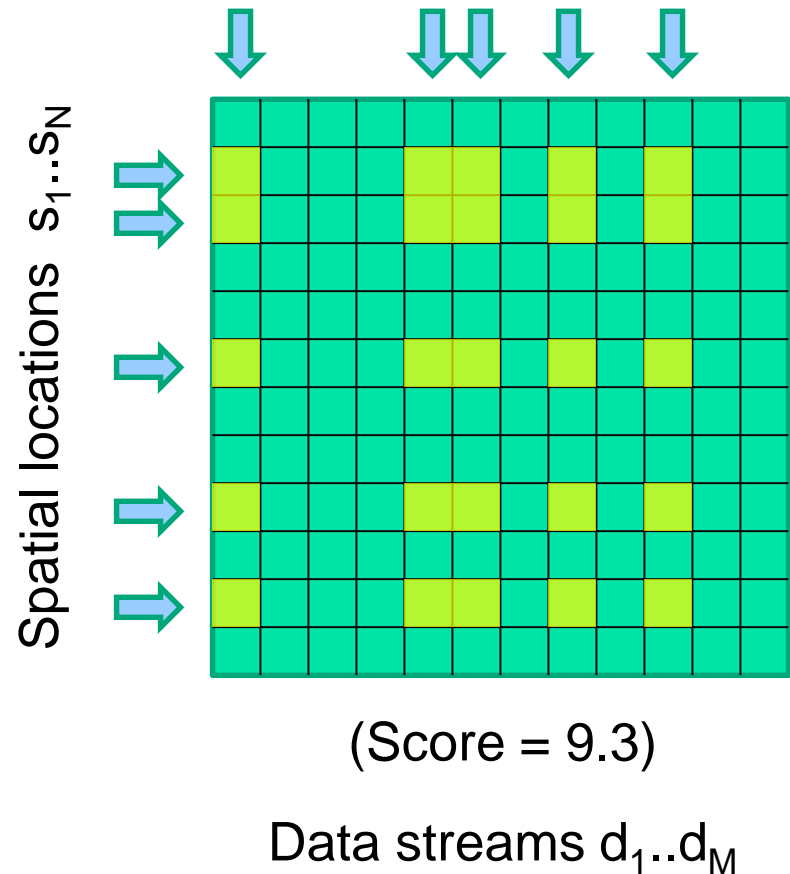
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.
- But it also allows us to efficiently optimize over subsets of **streams** for a given subset of **locations**...
- So we can jointly optimize over subsets of streams **and** locations by iterating between these steps!



Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

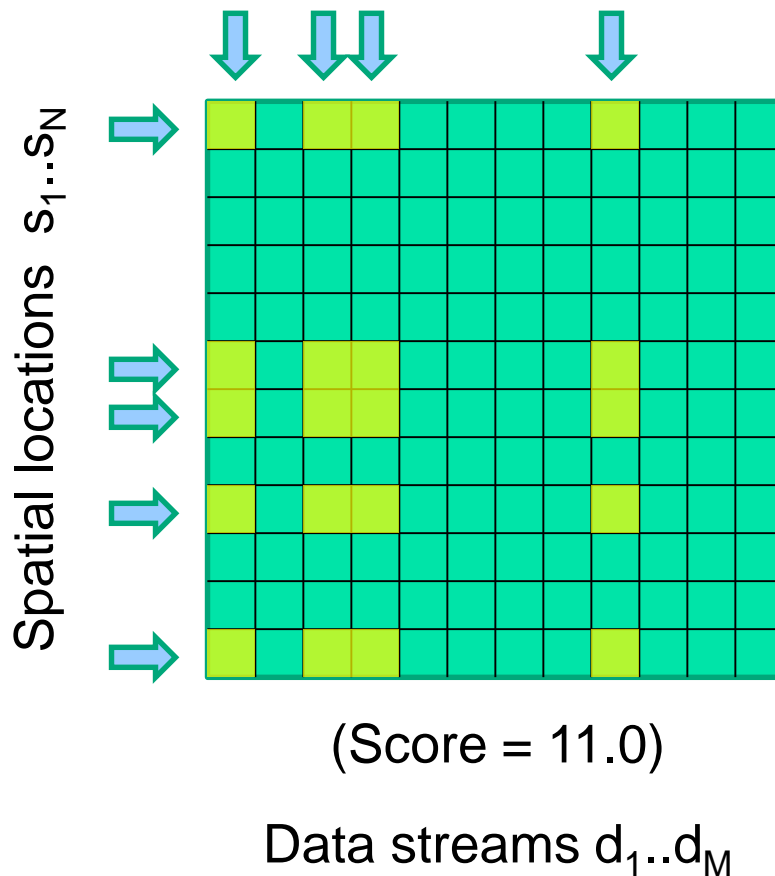
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.
- But it also allows us to efficiently optimize over subsets of **streams** for a given subset of **locations**...
- So we can jointly optimize over subsets of streams **and** locations by iterating between these steps!



Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.
- But it also allows us to efficiently optimize over subsets of **streams** for a given subset of **locations**...
- So we can jointly optimize over subsets of streams **and** locations by iterating between these steps!
- Converges to local maximum: we do multiple random restarts to approach the global maximum.
- For general datasets, a similar approach* can be used to jointly optimize over subsets of data records and attributes.

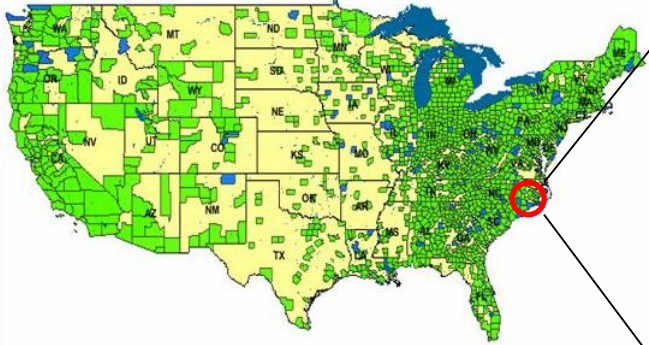


*McFowland, Speakman, and Neill, *JMLR*, 2013

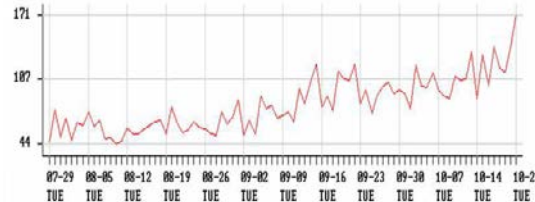
Outline of this talk

- Subset scanning for pattern detection
- **Multidimensional subset scan**
- Application #1: Event detection
 - outbreak detection, drug overdose surveillance
- Application #2: Discovery of heterogeneous treatment effects from observational data
 - patterns of patient care that impact outcomes
- Application #3: Auditing black-box classifiers to discover systematic biases
 - bias in criminal justice recidivism risk prediction

Multivariate event detection



Spatial time series data from spatial locations s_i (e.g. zip codes)



Time series of counts $c_{i,m}^t$ for each zip code s_i for each data stream d_m .

Outbreak detection

- d_1 = respiratory ED
- d_2 = constitutional ED
- d_3 = OTC cough/cold
- d_4 = OTC anti-fever
(etc.)

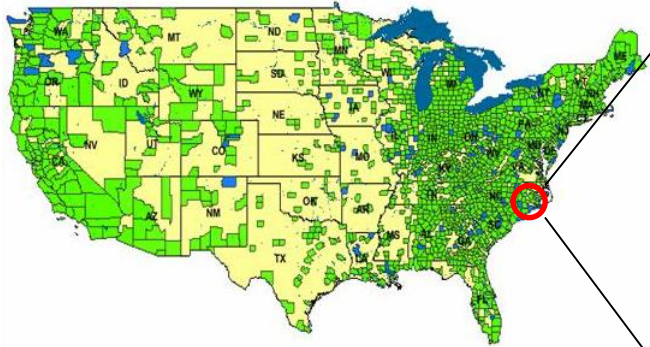
Main goals:

- Detect** any emerging events.
- Pinpoint** the affected subset of locations and time duration.
- Characterize** the event, e.g., by identifying the affected streams.

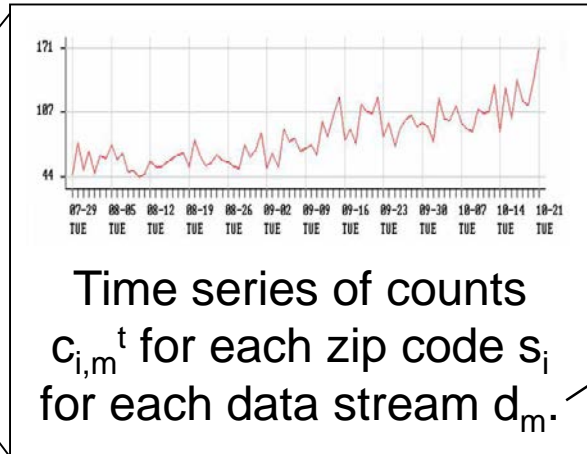
Compare hypotheses:

- $H_1(D, S, W)$
- D = subset of streams
- S = subset of locations
- W = time duration
- vs. H_0 : no events occurring

Multidimensional event detection



Spatial time series data from spatial locations s_i (e.g. zip codes)



Time series of counts $c_{i,m}^t$ for each zip code s_i for each data stream d_m .

Outbreak detection

- d_1 = respiratory ED
- d_2 = constitutional ED
- d_3 = OTC cough/cold
- d_4 = OTC anti-fever (etc.)

Additional goal: identify any differentially affected **subpopulations** P of the monitored population.

- Gender (male, female, both)
- Age groups (children, adults, elderly)
- Ethnic or socio-economic groups
- Risk behaviors: e.g. intravenous drug use, multiple sexual partners

More generally, assume that we have a set of additional discrete-valued attributes $A_1..A_J$ observed for each individual case. We identify not only the affected streams, locations, and time window, but also a **subset** of values for each attribute.

Multidimensional subset scan

- Our **MD-Scan** framework (Neill & Kumar, 2013) extends LTSS to the multidimensional case:
 - For each time window and spatial neighborhood (center + k-nearest neighbors), we do the following:
 1. Start with randomly chosen subsets of **locations** S , **streams** D , and **values** V_j for each attribute A_j ($j=1..J$).
 2. Choose an attribute A (randomly or sequentially) and use LTSS to find the highest scoring subset of values, locations, or streams, conditioned on all other attributes.
***** Linear rather than exponential in arity of A *****
 3. Iterate step 2 until convergence to a local maximum of the score function $F(D, S, W, \{V_j\})$, and use multiple restarts to approach the global maximum.

Outline of this talk

- Subset scanning for pattern detection
- Multidimensional subset scan
- **Application #1: Event detection**
 - outbreak detection, drug overdose surveillance
- Application #2: Discovery of heterogeneous treatment effects from observational data
 - patterns of patient care that impact outcomes
- Application #3: Auditing black-box classifiers to discover systematic biases
 - bias in criminal justice recidivism risk prediction

MD-Scan for event detection

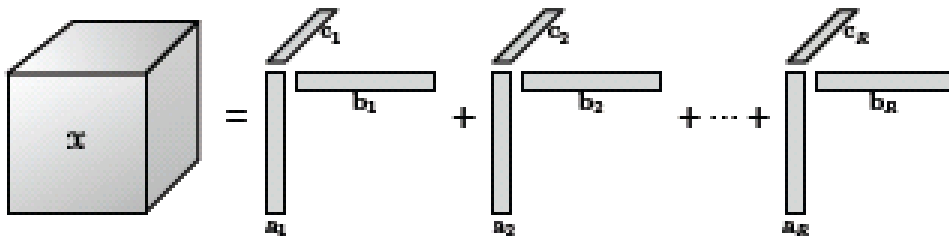
- As in the multivariate scan case, we can aggregate counts and baselines and maximize the EBP scan statistic over subsets. But how to get baselines?
- Original approach: compute separate baselines for each tensor cell (e.g., by 28-day moving average).
 - Statistical challenge: data sparsity leads to increasingly poor baseline estimates.
 - Computational challenge: very large tensor, often with dozens of modes, so need sparse representation.
 - We don't really believe that any baselines are zero!
- Solution: tensor decomposition!
 - 1) How to efficiently decompose?
 - 2) How to efficiently compute baselines?

Efficient factorization

- PARAFAC decomposition: approximate tensor by sum of outer products,

$$X = \sum_{r=1..R} (a^{(r)} \circ b^{(r)} \circ c^{(r)} \circ \dots)$$

or equivalently, $x_{ijk\dots} = \sum_{r=1..R} (a_i^{(r)} b_j^{(r)} c_k^{(r)} \dots)$



vectors = $R * \#$ modes

Each vector is of length = arity of that mode (or # of values of that attribute).

- Very large, sparse, high-order tensors: we want to run in time proportional to # of non-zero elements and independent of tensor size (product of arities).

Tensor power method

- Partial solution: fast rank-1 tensor decomposition.

Algorithm 2 Tensor Power Method

1. Initialize $\hat{X} = X$.
 2. For $k = 1 \dots K$
 - (a) Repeat until converge:
 - i. $u_k \leftarrow \hat{X} \times_2 v_k \times_3 w_k / \|\hat{X} \times_2 v_k \times_3 w_k\|_2$.
 - ii. $v_k \leftarrow \hat{X} \times_1 u_k \times_3 w_k / \|\hat{X} \times_1 u_k \times_3 w_k\|_2$.
 - iii. $w_k \leftarrow \hat{X} \times_1 u_k \times_2 v_k / \|\hat{X} \times_1 u_k \times_2 v_k\|_2$.
 - (b) $d_k \leftarrow \hat{X} \times_1 u_k \times_2 v_k \times_3 w_k$.
 - (c) $\hat{X} \leftarrow \hat{X} - d_k u_k v_k w_k$.
-

Compute successive rank-1 components by block coordinate-wise computation, subtract out, repeat on residuals.

- Easy to apply to sparse data. For example:
 - Step (a)(i). Zero u , then for each data point (i, j, k, value) add $\text{value} * v_j * w_k$ to u_i , then normalize u .
 - Step (b). Zero d , then for each data point (i, j, k, value) add $\text{value} * u_i * v_j * w_k$ to d .
- Good news: linear in # non-zeros of X .
- Bad news: for successive components after the first, X is no longer sparse!

Improved tensor power method

- Partial solution: fast rank-1 tensor decomposition.

Algorithm 2 Tensor Power Method

1. Initialize $\hat{X} = X$.
 2. For $k = 1 \dots K$
 - (a) Repeat until converge:
 - i. $u_k \leftarrow \hat{X} \times_2 v_k \times_3 w_k / \|\hat{X} \times_2 v_k \times_3 w_k\|_2$.
 - ii. $v_k \leftarrow \hat{X} \times_1 u_k \times_3 w_k / \|\hat{X} \times_1 u_k \times_3 w_k\|_2$.
 - iii. $w_k \leftarrow \hat{X} \times_1 u_k \times_2 v_k / \|\hat{X} \times_1 u_k \times_2 v_k\|_2$.
 - (b) $d_k \leftarrow \hat{X} \times_1 u_k \times_2 v_k \times_3 w_k$.
 - (c) $\hat{X} \leftarrow \hat{X} - d_k u_k v_k w_k$.
-

Compute successive rank-1 components by block coordinate-wise computation, subtract out, repeat on residuals.

- Do not modify X , but change update steps to take previous components into account. For example:
 - Step (a)(i). **Initialize** u , then for each data point (i, j, k, value) add $\text{value} \cdot v_j \cdot w_k$ to u_i , then normalize u .
 - Initialization: Zero $u^{(r)}$, then for each previous component $j=1..r-1$, subtract $\psi^{(j)} u^{(j)}$ from $u^{(r)}$, where $\psi^{(j)} = (v^{(j)} \cdot v^{(r)}) (w^{(j)} \cdot w^{(r)})$.
- Now X remains sparse, and we remain independent of tensor size for arbitrary # of PARAFAC components.

Computing baselines

- Given PARAFAC representation, the aggregate baseline of subset $S = S_1 \times S_2 \times \dots \times S_M$ is:

$$B = \sum_{r=1..R} \prod_{m=1..M} \sum_{i \in S_m} u_{i,m}^{(r)},$$

where $u_{i,m}^{(r)}$ is the i^{th} value of the m^{th} -mode vector of the r^{th} PARAFAC component.

- Example of why this works, for three modes:

$$\begin{aligned} B &= \sum_{i \in S_1} \sum_{j \in S_2} \sum_{k \in S_3} b_{ijk} \\ &= \sum_{i \in S_1} \sum_{j \in S_2} \sum_{k \in S_3} \sum_{r=1..R} u_i^{(r)} v_j^{(r)} w_k^{(r)} \\ &= \sum_{r=1..R} \left(\sum_{i \in S_1} u_i^{(r)} \right) \left(\sum_{j \in S_2} v_j^{(r)} \right) \left(\sum_{k \in S_3} w_k^{(r)} \right) \end{aligned}$$

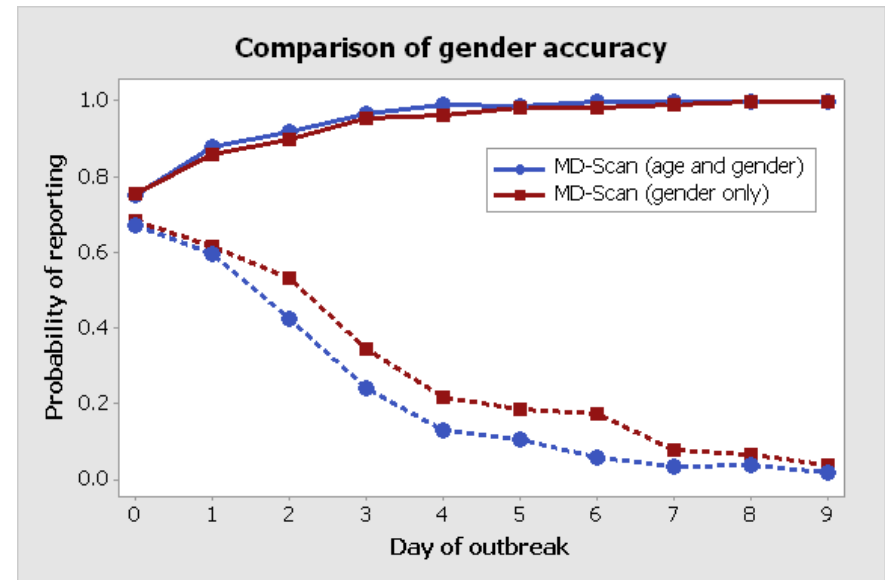
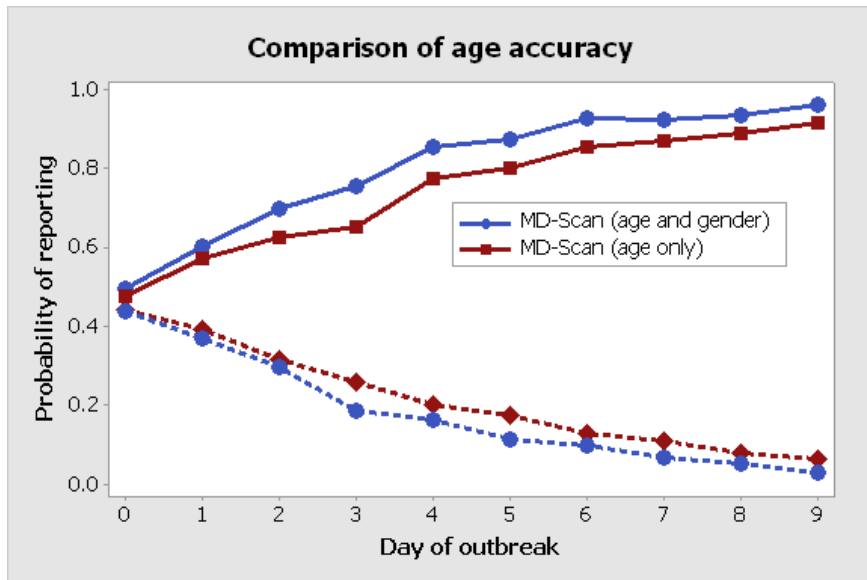
- By writing the sum of products as a product of sums, we can compute in time proportional to $|S_1| + |S_2| + \dots + |S_M|$ rather than $|S_1| \times |S_2| \times \dots \times |S_M|$.

Evaluation of MD-Scan

- We first evaluated the detection performance of MD-Scan for detecting simulated disease outbreaks injected into real-world Emergency Department data from Allegheny County, PA.
- For outbreaks with differential effects by age and gender, MD-Scan demonstrated **more timely** and **more accurate** detection, and accurately **characterized** the affected subpopulations.

1) Identifying affected subpopulations

By the midpoint of the outbreak, MD-Scan is able to correctly identify the affected gender and age deciles with high probability, without reporting unaffected subpopulations.



Proportions of correct and incorrect groups reported vs. time since start of outbreak.

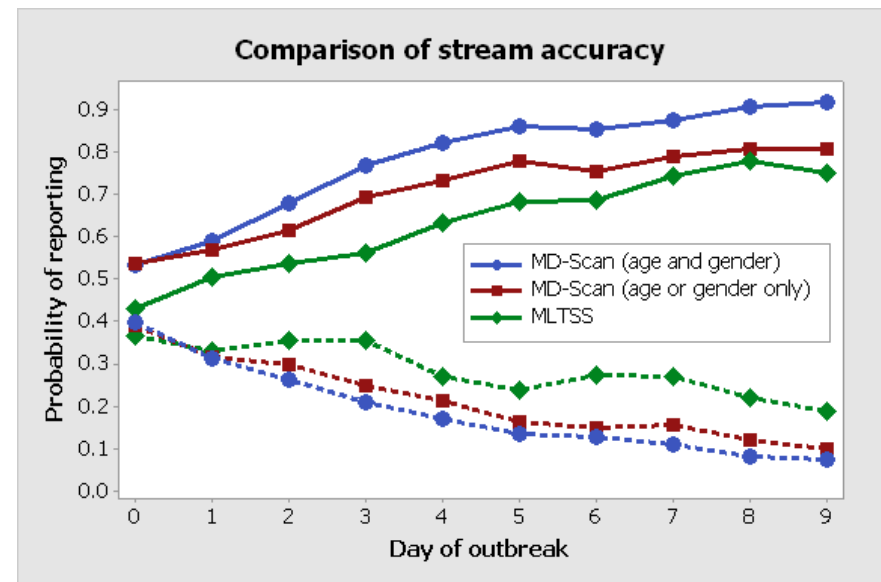
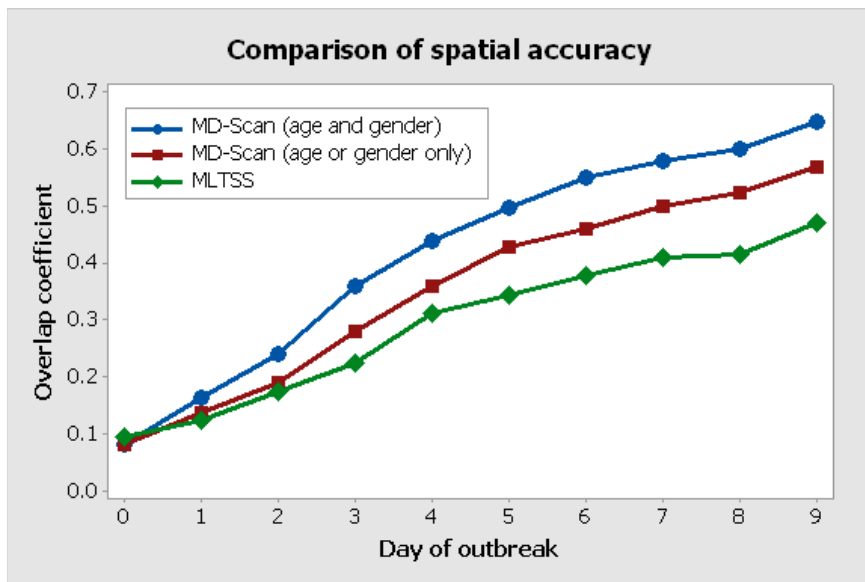
Solid lines: affected gender and/or age deciles. Dashed lines: unaffected.

Blue lines: outbreaks with differential effects by both age and gender (easier).

Red lines: outbreaks with differential effects by age or gender only (harder).

2) Characterizing affected streams

As compared to the previous state of the art (multivariate linear-time subset scanning), MD-Scan is better able to characterize the affected spatial locations and subset of the monitored streams.



Left: overlap coefficient between true and detected subsets of spatial locations.
Right: Proportions of correct and incorrect streams reported vs. day of outbreak.

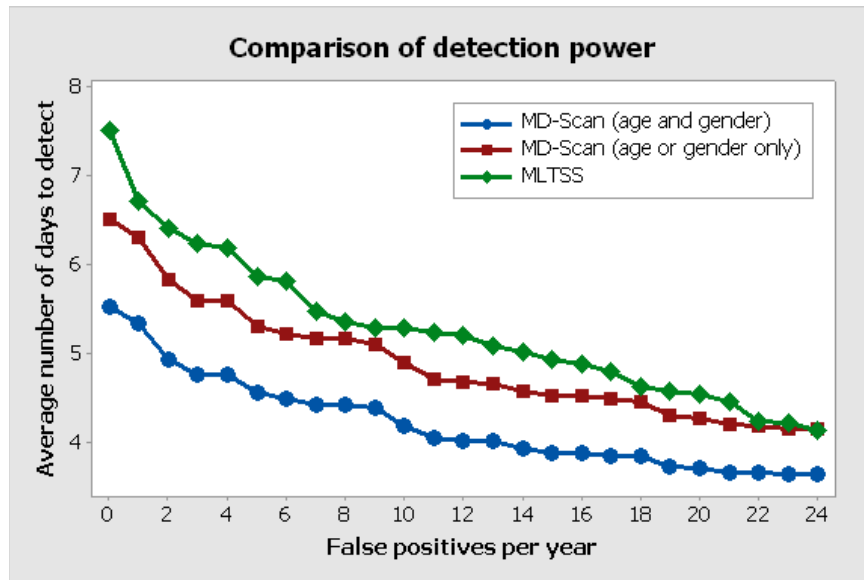
Blue lines: outbreaks with differential effects by both age and gender (easier).

Red lines: outbreaks with differential effects by age or gender only (harder).

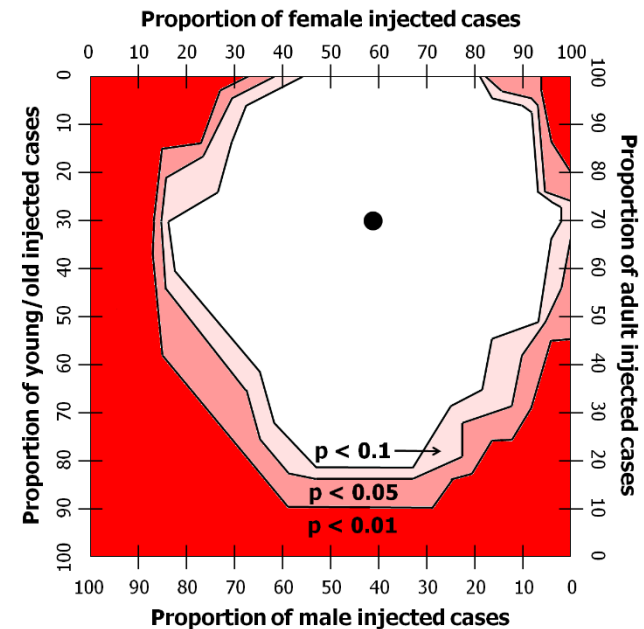
Green lines: MLTSS, ignoring age and gender information

3) Timeliness of outbreak detection

MD-Scan achieved significantly more timely detection for outbreaks that were sufficiently biased by age and/or gender.



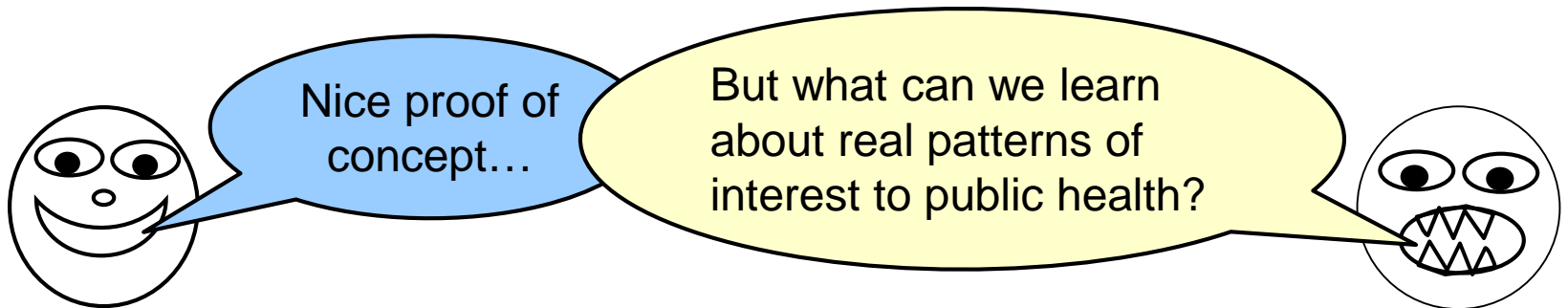
For outbreaks with strong age and gender biases, time to detection improved from 5.2 to 4.0 days at a fixed false positive rate of 1/month.



Smaller biases in age or gender were sufficient for significant improvements; even when no age/gender signal is present, MD-Scan performs comparably to MLTSS.

Evaluation of MD-Scan

- We first evaluated the detection performance of MD-Scan for detecting simulated disease outbreaks injected into real-world Emergency Department data from Allegheny County, PA.
- For outbreaks with differential effects by age and gender, MD-Scan demonstrated **more timely** and **more accurate** detection, and accurately **characterized** the affected subpopulations.



Allegheny County Overdose Data

- We analyzed county medical examiner data for fatal accidental drug overdoses, 2008-2015.
- ~2000 cases: for each overdose victim, we have date, location (zip), age, gender, race, and the set of drugs present in their system.
- Reduced to 30 dimensions (age decile, gender, race, presence/absence of 27 common drugs) plus space and time.
- Clusters discovered by MD-Scan were shared with Allegheny County Dept. of Human Services.

MD-Scan Overdose Results (1)



Fentanyl is a dangerous drug which has been a huge problem in western PA.

It is often mixed with white powder heroin, or sold disguised as heroin.

January 16-25, 2014:

14 deaths county-wide from fentanyl-laced heroin.

March 27 to April 21, 2015:

26 deaths county-wide from fentanyl, heroin only present in 11.

January 10 to February 7, 2015:

Cluster of 11 fentanyl-related deaths, mainly black males over 58 years of age, centered in Pittsburgh's downtown Hill District.

Very unusual demographic: common dealer / shooting gallery?

Started in the SE suburbs of Pittsburgh, including a cluster of 5 cases around McKeesport between March 27 and April 8.

Cluster score became significant March 29th (4 nearby cases, white males ages 20-49) and continued to increase through April 20th.

Fentanyl, heroin, and combined deaths remained high through end of June (>100).

MD-Scan Overdose Results (2)

Another set of discovered overdose clusters each involved a combination of Methadone and Xanax.



The combination produces a strong high but can be deadly (~30% of methadone fatal ODs).

From 2008-2012: multiple M&X OD clusters, 3-7 cases each, localized in space and time.

Methadone: an opioid used for chronic pain relief and to treat heroin addiction, but also addictive and risk of OD.

From 2013-2015: no M&X overdose clusters; 33% and 47% drops in yearly methadone and M&X deaths respectively.



Why did these deaths cluster, when methadone and methadone + other benzo deaths did not?

Xanax (alprazolam): a benzodiazepine prescribed for panic and anxiety disorders.

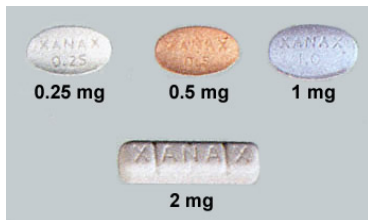
What factors could explain the dramatic reduction in M&X overdose clusters?

MD-Scan Overdose Results (2)

Another set of discovered overdose clusters each involved a combination of Methadone and Xanax.



Methadone: an opioid used for chronic pain relief and to treat heroin addiction, but also addictive and risk of OD.



Xanax (alprazolam): a benzodiazepine prescribed for panic and anxiety disorders.

Increased state oversight of methadone clinics and prescribing physicians after passage of the Methadone Death and Incident Review Act (Oct 2012).

Approval of generic suboxone (buprenorphine + naloxone) in early 2013 lowered cost of suboxone treatment as an alternative to methadone clinics.

Why did these deaths cluster, when methadone and methadone + other benzo deaths did not?

What factors could explain the dramatic reduction in M&X overdose clusters?

Outline of this talk

- Subset scanning for pattern detection
- Multidimensional subset scan
- Application #1: Event detection
 - outbreak detection, drug overdose surveillance
- **Application #2: Discovery of heterogeneous treatment effects from observational data**
 - patterns of patient care that impact outcomes
- Application #3: Auditing black-box classifiers to discover systematic biases
 - bias in criminal justice recidivism risk prediction

Case study #2: Discovering anomalous patterns of care

Joint work with Sriram Somanchi and Edward McFowland III

- Given health insurance claims data, we wish to identify a **treatment** and corresponding **sub-population** for whom that treatment leads to significantly better or worse outcomes.
 - Observational data; multiple treatments.
 - Population characteristics vary on multiple dimensions.
 - Identify **most significant** combinations of treatment and sub-population.

“For males over 50 with congestive heart failure and certain co-morbidities, taking Carvidilol is associated with longer stay in hospital.”
– Patrick, EPD Lab healthcare analyst (after significant manual effort)

Problem formulation

- Let $X = (X_1, X_2, \dots, X_N)$ be the set of observed covariates for a patient (demographics, diagnoses, etc.)
- Let T_1, T_2, \dots, T_M be the set of available treatments.
- Let Y be the scalar outcome of interest (for example, number of hospitalizations in some time period following treatment).

Our goals

- Estimate the distribution of potential outcomes, for treatment assignments $T_j = 1$ and $T_j = 0$ respectively, for any given subpopulation S :

$$f_{j1,S} = f(y^{(1)} \mid x \in S)$$

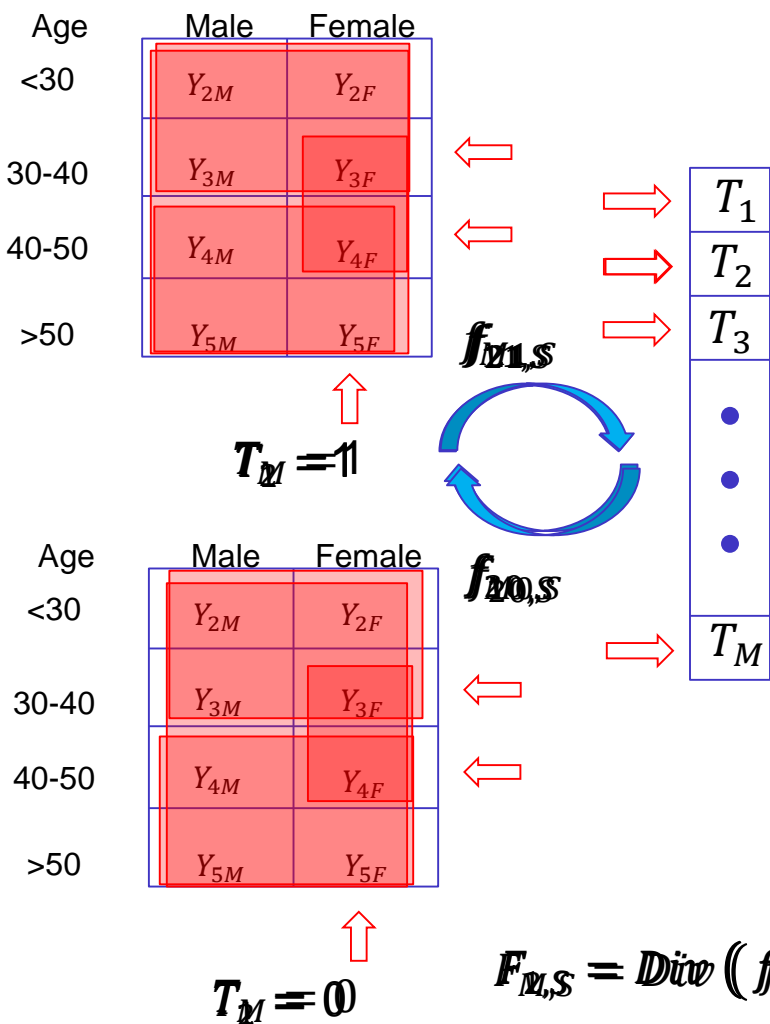
$$f_{j0,S} = f(y^{(0)} \mid x \in S)$$

- Find the combination of treatment and subpopulation that maximizes some measure of divergence, $Div(f_{j1,S}, f_{j0,S})$.

Age	Male	Female
<30	Y_{2M}	Y_{2F}
30-40	Y_{3M}	Y_{3F}
40-50	Y_{4M}	Y_{4F}
>50	Y_{5M}	Y_{5F}

T_1
T_2
T_3
•
•
•
T_M

Anomalous Patterns of Care Scan



1. Start with a random sub-population S
2. For each T_j
 - a. Compute propensity scores
 - b. Reweight outcome distributions
 - c. Compute divergence $F_{j,S}$
3. Choose treatment: $j^* = \text{argmax}_j F_{j,S}$
4. Reweight entire population outcomes based on T_{j^*}
5. Use MD-Scan to identify $S^* = \text{argmax}_S F_{j^*,S}$
6. Set $S = S^*$ and repeat steps 2 to 5 until score stops increasing
7. Repeat steps 1-6 for R times
8. Compute statistical significance by randomization testing

Iterative Ascent algorithm between sub-populations and treatments

Challenges for APC-Scan

- We use the **expectation-based Poisson** scan statistic, scanning over the treatment individuals. Each has an observed count (number of visits) and an expected count estimated from the control individuals.
- Challenge 1: data sparsity. May be few or no controls who match the treated individual.
 - Solution: learn a predictive model for $y | x$ from control individuals, then use to predict y for each treatment individual.

Challenges for APC-Scan

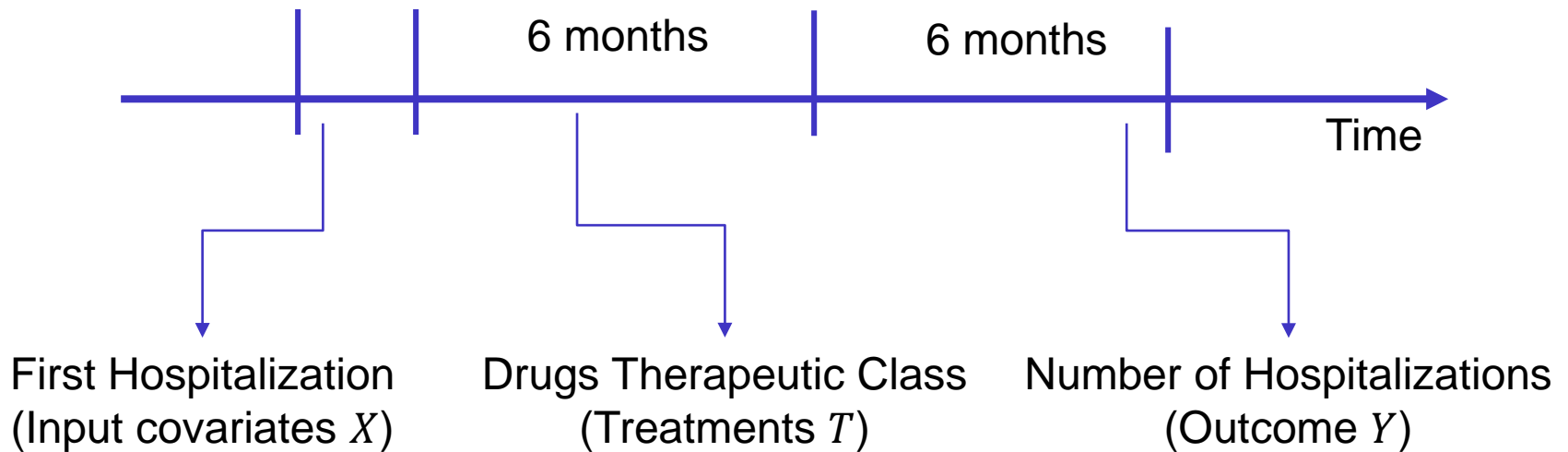
- We use the **expectation-based Poisson** scan statistic, scanning over the treatment individuals. Each has an observed count (number of visits) and an expected count estimated from the control individuals.
- Challenge 2: **selection into treatment**.
A treatment could have worse outcomes just because it is typically given to sicker patients.
 - Partial solution: use inverse propensity score weighting to account for observable differences between treatments and controls.

Inverse propensity score weighting

- We are estimating average treatment effect on the treated (ATT), which is a weighted average of the conditional average treatment effects weighted by probability of treatment.
- Control individuals are weighted by $\frac{p}{1-p}$, where the estimated treatment probability $p = \Pr(T_j = 1 \mid x)$.
- We learn baselines for each treated individual using the weighted control data, then scan over the (unweighted) treated individuals.
- This produces an unbiased estimate of ATT if unconfoundedness holds, i.e., if $\{y^{(0)}, y^{(1)}\} \perp T_j \mid x$.

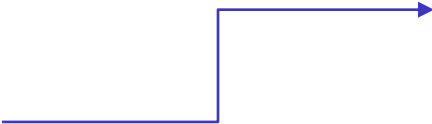
Highmark claims data

- ~125K patients with primary or admission diagnosis as “diseases of the circulatory system” during 2008-2014.



Highmark claims data

- Covariates (X) included:
 - Demographics
 - Median income in patient's home zip code
 - Diagnosis (primary and secondary)
 - Charlson Comorbidity Index
 - Length of current stay
 - Previous outpatient visits
- Treatments (T_j)
 - Drug Therapeutic Class
- Outcome (Y)
 - Number of hospitalizations



Bronchial Dilators
Glucocorticoids
Thyroid Preparations
Diabetic Therapy
Lipotropics
Hypotensives
Vasodilators
Digitalis Preparations
Cardiovascular
Preparations
Anticoagulants
Diuretics

Highest scoring detected pattern

Glucocorticoids significantly increase mean number of hospitalizations following treatment in the subpopulation of hypertensive, overweight/obese males with endocrine disorders.

- Identified subpopulation characteristics (N = 1,977):
 - Gender = Male
 - Hypertension = Yes
 - Diabetes = Yes or No
 - BMI = Obese or Overweight
 - Age Ranges = 40-60 or 60-80
 - Primary diagnosis = Ischemic heart disease, Heart failure, or Cerebrovascular heart disease.
 - Secondary diagnosis = Endocrine

	Glucocorticoids	
	Yes	No
Number of Patients	264	1713
Mean Number of Hospitalizations	0.606 (0.069)	0.280 (0.016)

Validation of our results

- There is a growing literature in the medical community relating glucocorticoids with cardiovascular issues:
 - Association using 10 years of observational data (Heart, 2004)
 - Metabolic and tissue level effects in heart (European Journal of Endocrinology, 2007)
 - Experiments at micro level analysis of glucocorticoids signaling certain receptors in heart for mice (J. Biochem. Molec. Biol., 2015)
- But no results on heterogeneity of effect across subpopulations!

Regression analysis

- We randomly split the data into:
 - 60% for running our APC Scan
 - 40% for running the regression analysis
- Regression with outcome Y as number of hospitalizations with Glucocorticoids as one of the independent variables X , for:
 - The entire population
 - The entire population with a dummy for the subpopulation identified by APC Scan
 - The subpopulation identified by APC Scan
 - The complementary subpopulation

Regression analysis (Poisson) on held-out data

	Number of Hospitalizations		Number of Hospitalizations		
	(1)	(2)	(3)	(4)	
Glucocorticoids	0.101*** (0.007)	0.099*** (0.007)	0.410*** (0.089)	0.099*** (0.007)	(1) Entire Population
Glucocorticoids* Subpopulation		0.265*** (0.088)			(2) Entire Population with dummy for the subpopulation
Subpopulation		-0.313*** (0.068)			(3) Subpopulation identified by APC- Scan
Age	0.079*** (0.004)	0.079*** (0.004)	-0.040 (0.079)	0.080*** (0.004)	(4) Remaining subpopulation not identified by APC-Scan
Females	0.116*** (0.008)	0.113*** (0.008)		0.113*** (0.008)	
Hypertensive	-0.163*** (0.008)	-0.161*** (0.008)		-0.161*** (0.008)	
Diabetic	0.286*** (0.008)	0.286*** (0.008)	0.193*** (0.089)	0.287*** (0.008)	
Obesity	0.007 (0.013)	0.020 (0.013)		0.020 (0.013)	
...	
Constant	-0.773*** (0.044)	-0.772*** (0.044)	-1.634*** (0.120)	-0.772*** (0.044)	
Observations	49,658	49,658	796	48,862	

10.6%

50.6%

*p<0.1; **p<0.05; ***p<0.01

Outline of this talk

- Subset scanning for pattern detection
- Multidimensional subset scan
- Application #1: Event detection
 - outbreak detection, drug overdose surveillance
- Application #2: Discovery of heterogeneous treatment effects from observational data
 - patterns of patient care that impact outcomes
- **Application #3: Auditing black-box classifiers to discover systematic biases**
 - bias in criminal justice recidivism risk prediction



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Source:
Julia Angwin,
Jeff Larson,
Surya Mattu and
Lauren Kirchner, *ProPublica*

Machine Bias

There's software used across the country to predict future criminals.
And it's biased against blacks.

Two Drug Possession Arrests



DYLAN FUGETT

Prior Offense
1 attempted burglary

Subsequent Offenses
3 drug possessions

LOW RISK

3



BERNARD PARKER

Prior Offense
1 resisting arrest without
violence

Subsequent Offenses
None

HIGH RISK

10

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

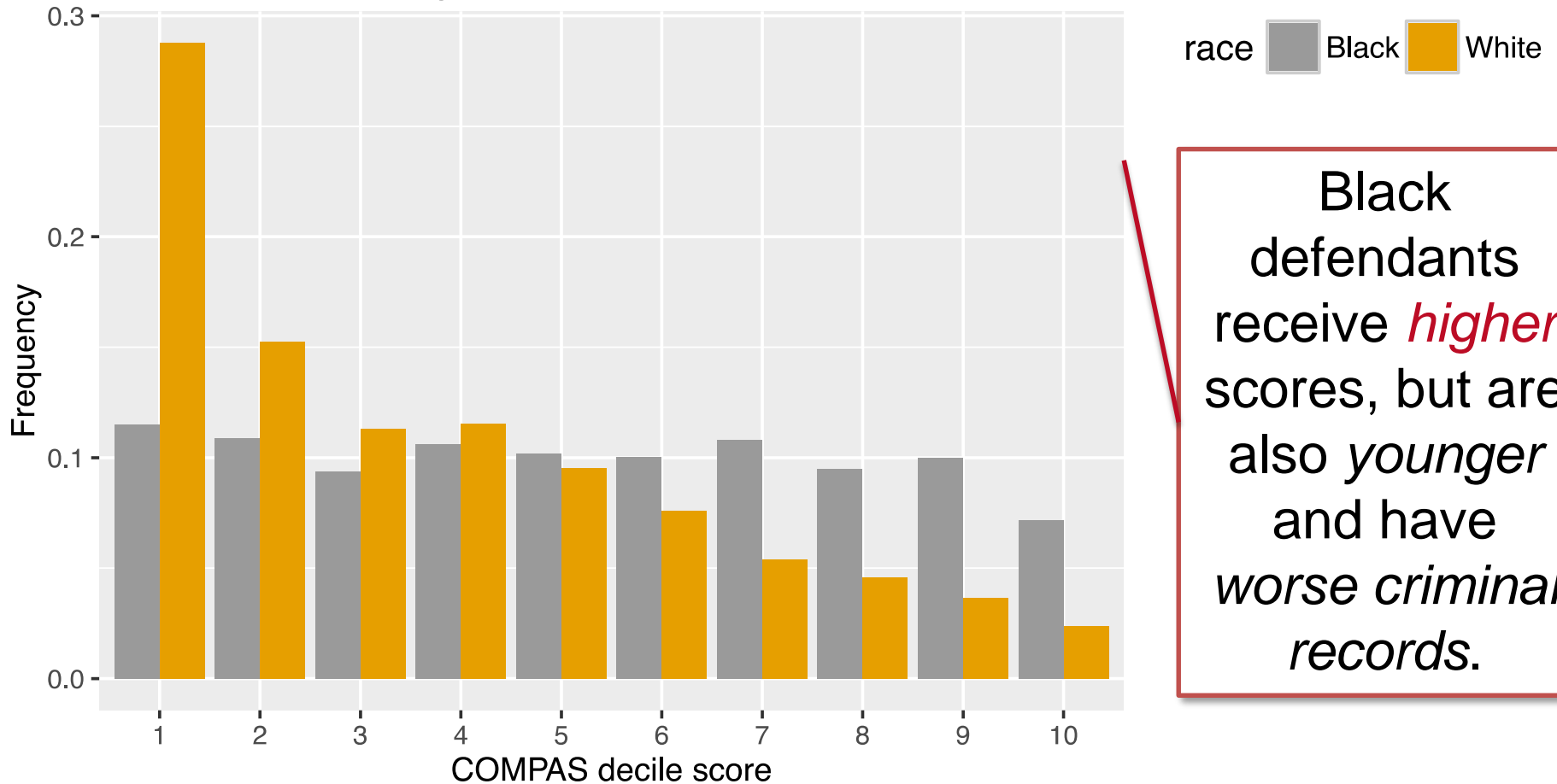
Broward County data

- Source: ProPublica's data on criminal defendants in Broward County, FL, in 2013-2014
- Outcome: re-arrests (!) assessed through April 2016.
- Score: **COMPAS** score from 1 (low risk) to 10 (high risk)

Background	Black (<i>n</i> = 3696)		White (<i>n</i> = 2454)
Age	32.7 (10.9)	<	37.7 (12.8)
Male (%)	82.4	>	76.9
Number of Priors	4.44 (5.58)	>	2.59 (3.8)
Any priors? (%)	76.4	>	65.9
Felony (%)	68.9	>	60.3
COMPAS Score	5.37 (2.83)	>	3.74 (2.6)

Sample averages (standard deviations)

Histograms of COMPAS scores

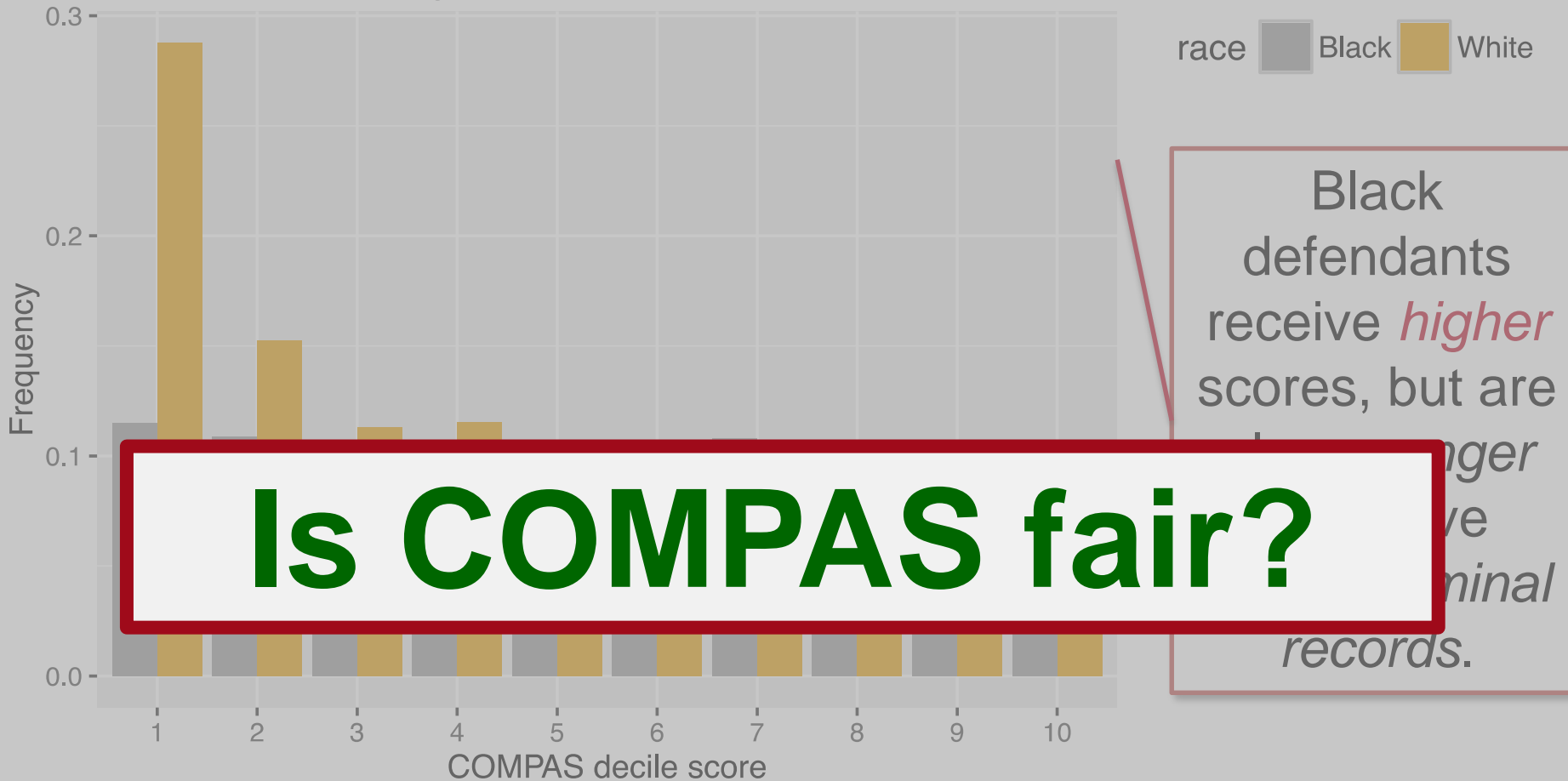


Black defendants receive *higher* scores, but are also *younger* and have *worse criminal records*.

Outcome	Black	White
Recidivism (%)	51.4	39.4
Violent Recidivism (%)	13.40	9.05

Observed recidivism is *higher* among Black defendants.

Histograms of COMPAS scores



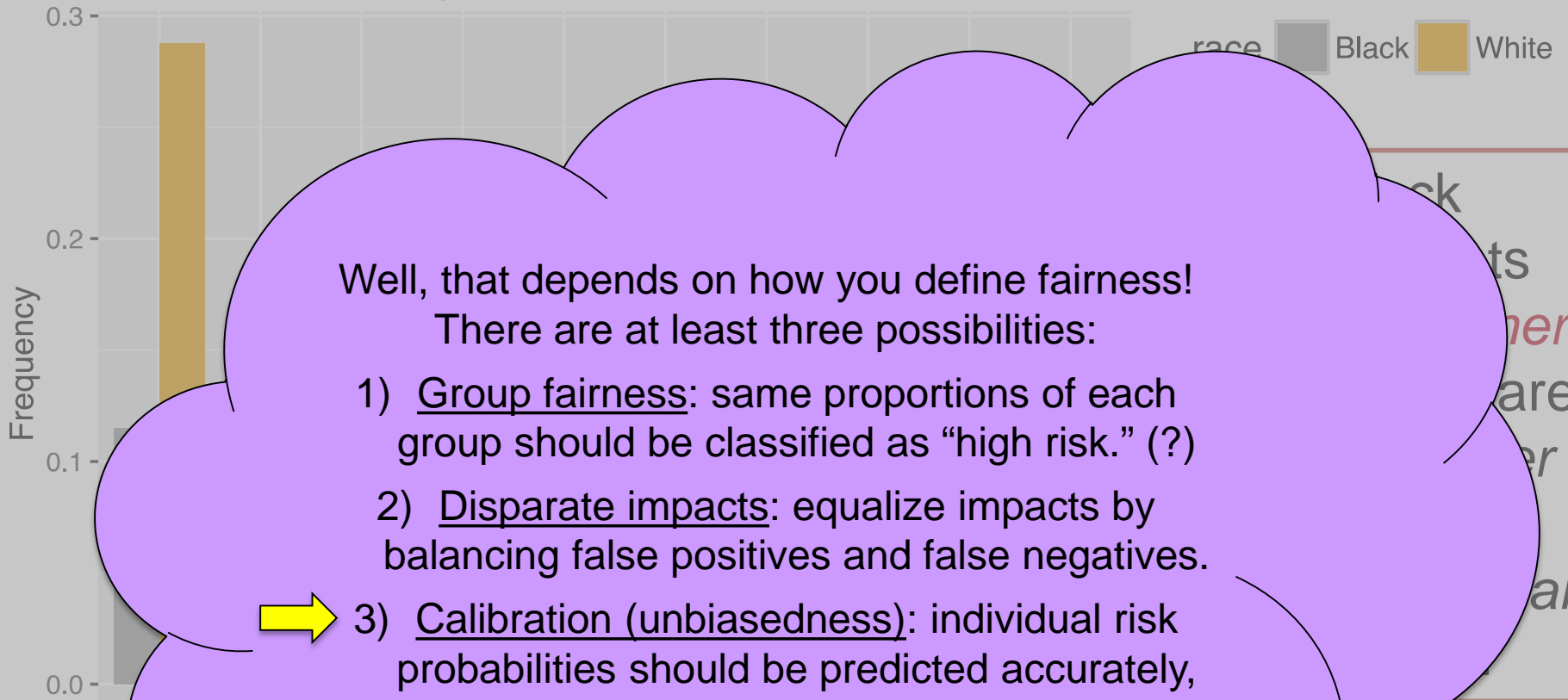
Is COMPAS fair?

Black defendants receive *higher* scores, but are *more likely to have criminal records.*

Observed recidivism is *higher* among Black defendants.

Outcome	Black	White
Recidivism (%)	51.4	39.4
Violent Recidivism (%)	13.40	9.05

Histograms of COMPAS scores



Well, that depends on how you define fairness!
There are at least three possibilities:

- 1) Group fairness: same proportions of each group should be classified as “high risk.” (?)
- 2) Disparate impacts: equalize impacts by balancing false positives and false negatives.
- 3) Calibration (unbiasedness): individual risk probabilities should be predicted accurately, **without systematic biases** based on race or any other combinations of attributes.



Outcome

Recidivism (%)	51.4	51.4
Violent Recidivism (%)	13.40	9.05

Recidivism is *higher* among Black defendants.

Bias scan (Zhang and Neill, 2016)

Our goal is to **detect** and **correct** any **systematic biases** in risk prediction that a classifier may have (i.e., over-predicting or under-predicting risk for a specific attribute or combination of attributes).

We developed a new variant of the multidimensional subset scan to identify subgroups where classifier predictions are significantly biased.

Assume a dataset with inputs x_i , binary labels $y_i \in \{0, 1\}$, and the classifier's risk predictions $\hat{p}_i = \Pr(y_i = 1)$.

Search space: subspaces defined by a subset of values for each attribute (e.g., “white and Asian males under 25”)

Score function: a log-likelihood ratio statistic. H_0 : \hat{p}_i correctly calibrated; $H_1(S)$: constant multiplicative increase or decrease in odds of $y_i = 1$ for subspace S .

$$F(S) = \max_q \log \prod_{s_i \in S} \frac{\Pr\left(y_i \sim \text{Bernoulli}\left(\frac{q\hat{p}_i}{1 - \hat{p}_i + q\hat{p}_i}\right)\right)}{\Pr(y_i \sim \text{Bernoulli}(\hat{p}_i))}$$

Bias scan (Zhang and Neill, 2016)

Our goal is to **detect** and **correct** any **systematic biases** in risk prediction that a classifier may have (i.e., over-predicting or under-predicting risk for a specific attribute or combination of attributes).

We developed a new variant of the multidimensional subset scan to identify subgroups where classifier predictions are significantly biased.

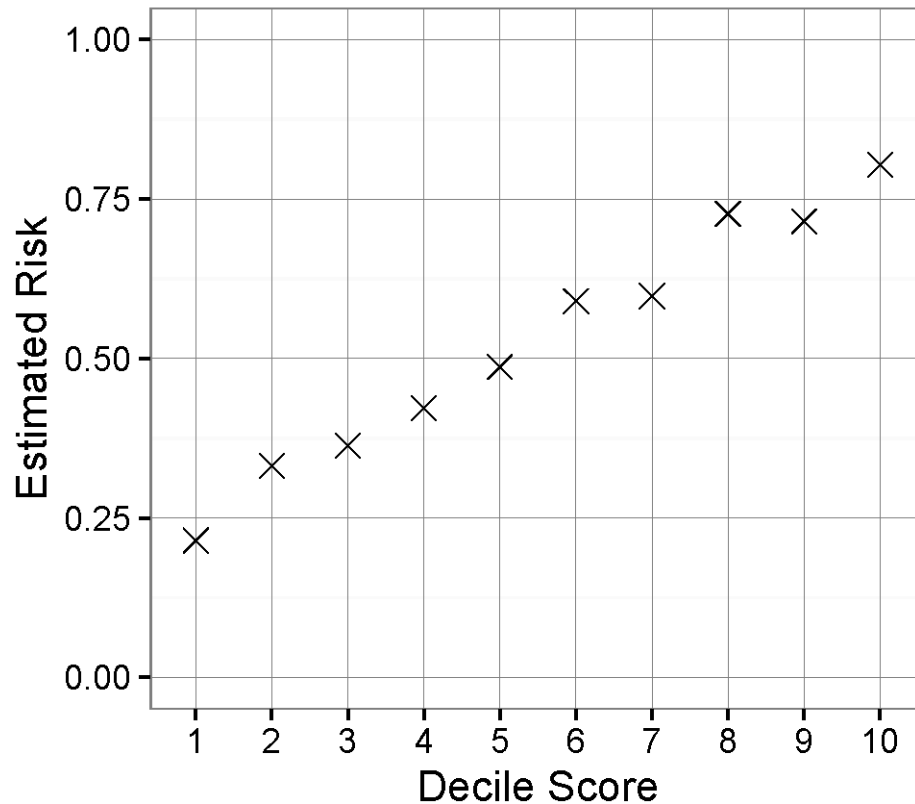
Assume a dataset with inputs x_i , binary labels $y_i \in \{0, 1\}$, and the classifier's risk predictions $\hat{p}_i = \Pr(y_i = 1)$.

Search space: subspaces defined by a subset of values for each attribute (e.g., “white and Asian males under 25”)

Score function: a log-likelihood ratio statistic. H_0 : \hat{p}_i correctly calibrated; $H_1(S)$: constant multiplicative increase or decrease in odds of $y_i = 1$ for subspace S .

For interpretability, we maximize the penalized score $F(S) - \log \prod |S_j|$, where attributes with no excluded values are ignored. For each conditional optimization, we can use the simple penalty, $\log(|S_j|) 1\{|S_j| < \text{arity}(A_j)\}$.

Results of bias scan on COMPAS

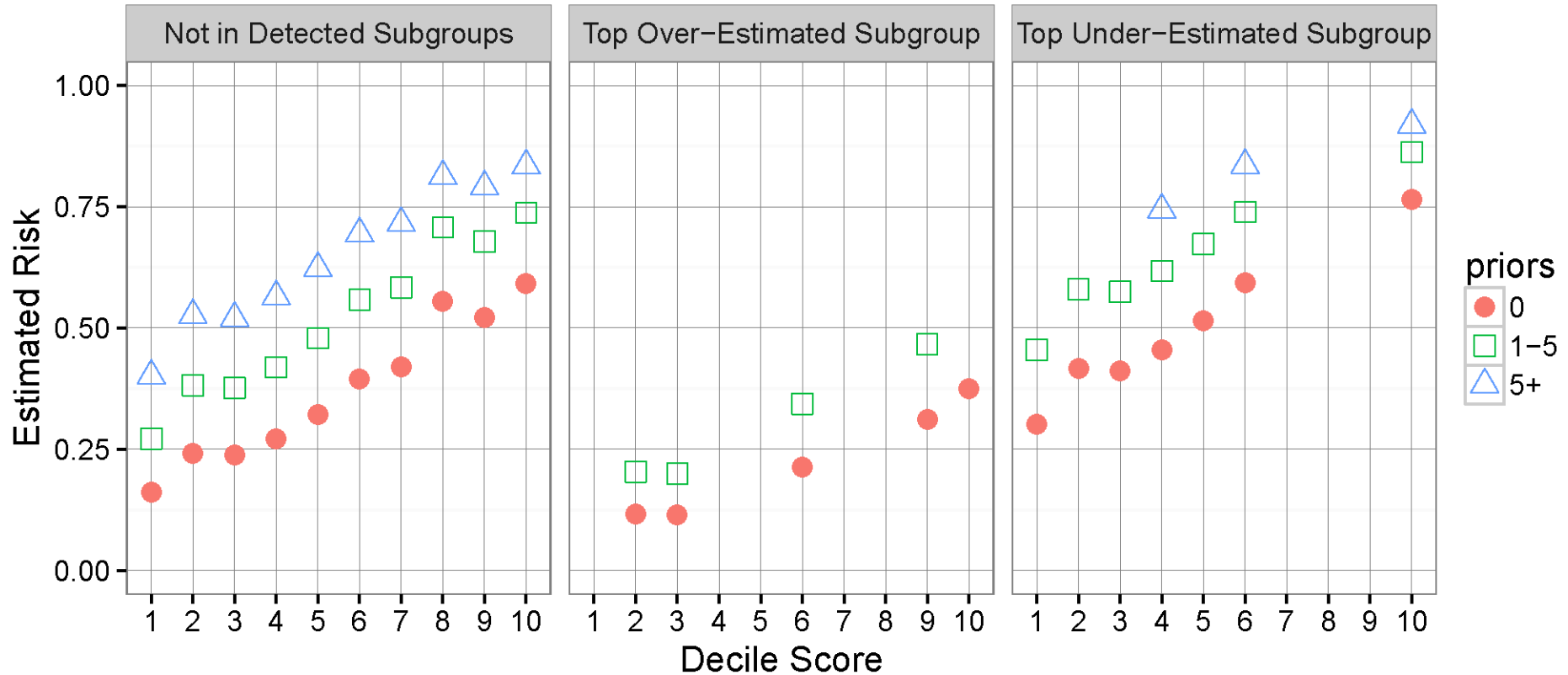


Start with maximum likelihood risk estimates for each COMPAS decile score.

Detection result 1: COMPAS underestimates the importance of prior offenses, overestimating risk for 0 priors, and underestimating risk for 5 or more priors.

Detection result 2: Even controlling for prior offenses, COMPAS still underestimates risk for males under 25, and overestimates risk for females who committed misdemeanors.

Results of bias scan on COMPAS



After controlling for prior offenses and membership in the two detected subgroups, there are no significant systematic biases in prediction.

Thorny question: given individual risk predictions, what should we do with them (e.g., how to avoid disparate impacts)?

Conclusions

Real-world problems at the societal scale require new computational methods to deal with both the **size** and the **complexity** of data.

Fast multidimensional subset scanning can serve as a fundamental building block for scalable pattern detection in massive, complex data.

With slight extensions, the same multidimensional scan framework can be used effectively across a variety of problems ranging from **event detection** to **causal inference** to **algorithmic fairness**.

Potential benefits to the **public good** include more timely detection of emerging outbreaks and trends in drug overdoses, improved patient outcomes, and fairer use of algorithms in criminal justice.

Acknowledgements

- Students and collaborators: Skyler Speakman, Ed McFowland, Sriram Somanchi, Tarun Kumar, Patrick Wedgeworth, William Herlands, Zhe Zhang
- Alexandra Chouldechova (background slides on COMPAS/ProPublica debate)
- Sriram Somanchi (APC-Scan slides).
- Funding support: NSF and MacArthur Foundation.

References

- D.B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society (Series B: Statistical Methodology)* 74(2): 337-360, 2012.
- D.B. Neill, E. McFowland III, and H. Zheng. Fast subset scan for multivariate event detection. *Statistics in Medicine* 32: 2185-2208, 2013.
- E. McFowland III, S. Speakman, and D.B. Neill. Fast generalized subset scan for anomalous pattern detection. *Journal of Machine Learning Research* 14: 1533-1561, 2013.
- S. Speakman, S. Somanchi, E. McFowland III, D.B. Neill. Penalized fast subset scanning. *Journal of Computational and Graphical Statistics* 25: 382-404, 2016.
- T. Kumar and D.B. Neill. Fast tensor scan for event detection and characterization. Revised version in preparation.
- D.B. Neill and W. Herlands. Machine learning for drug overdose surveillance. *Proceedings of the Bloomberg Data for Good Exchange Conference*, 2017.
- S. Somanchi, E. McFowland III, and D.B. Neill. Detecting anomalous patterns of care using health insurance claims. In preparation.
- Z. Zhang and D.B. Neill. Identifying significant predictive bias in classifiers. <https://arxiv.org/pdf/1611.08292.pdf>. *Proceedings of the NIPS Workshop on Interpretable Machine Learning*, 2016.



Thanks for listening!

More details on our web site:

<http://epdlab.heinz.cmu.edu>

Or e-mail me at:

neill@cs.cmu.edu